Supporting Information

Evaluation of *In Silico* Multi-Feature Libraries for Providing Evidence for the Presence of Small Molecules in Synthetic Blinded Samples

Jamie Nuñez[1], Sean Colby[1], Dennis Thomas[1], Malak Tfaily[1], Nikola Tolic[1], Elin Ulrich[2], Jon Sobus[2], Thomas O. Metz[1,*], Justin Teeguarden,[1,3,*] Ryan Renslow[1,*]

[1]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

[2]U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, North Carolina 27709, United States

[3]Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, Oregon 97331, United States

* ryan.renslow@pnnl.gov

This document contains additional information to support the manuscript cited above. Included here are additional method details, captions for the data provided in SupportingData.xlsx, and additional figures that support claims made in the manuscript. Please see manuscript for context.

Contents

Methods

SI 1. Sample Preparation.

Ten mixtures and a blank in dimethylsulfoxide (DMSO) were received from the U.S. Environmental Protection Agency (EPA) and stored at -80 °C until analysis. Each mixture contained an unknown number of substances (later revealed as 95-365 substances), with all substances selected from EPA's ToxCast chemical library[43]. Further details on the EPA's Non-Targeted Analysis Collaborative Trial (ENTACT, Phase 1) inter-laboratory challenge are outlined in Sobus et al. (2017)[42] and Ulrich et al. (2019).[41]

For analyses using electrospray ionization (ESI), a 1:20 mixing ratio was used where 5 µL of each mixture sample and blank was diluted with 95 µL of methanol containing 0.5% acetic acid. For analyses using atmospheric pressure photoionization (APPI), only glass needles and vials were used to prepare the samples since non-polar molecules adhere to plastics. Similar to the ESI analyses, a 1:20 mixing ratio was used for all mixture samples and blank where 20 µL of each mixture was diluted in 340 µL MeOH and 40 µL acetone with 0.5% acetic acid.

SI 2. Ion-Mobility Spectrometry-Mass Spectrometry.

2.1 Instrumentation. All samples were analyzed with an Agilent 6560 drift tube ion mobility spectrometry, quadrupole time-of-flight mass spectrometer (DTIMS-QTOF MS)[1,2] using ESI and APPI in both positive and negative ionization modes. For ESI, 5 µL of each sample was injected into the DTIMS-QTOF MS ESI source at 300 nL/min for a 1-minute data acquisition. APPI source conditions were optimized as follows: gas temperature, 310 °C; drying gas, 2.5 L/min; nebulizer, 30 psig; VCap, 1100 V; and Vaporizer, 310 °C. The solvent composition used for APPI is slightly different than ESI since substances like toluene, acetone, anisole, and chlorobenzene are needed to increase APPI ionization efficiency, with 10% usually being optimal. 50 µL of each sample was injected into the DTIMS-QTOF MS APPI source at a flow rate of 40 µL/min for a 1-minute data acquisition.

The DTIMS instrument was outfitted with the commercial gas kit (Alternate Gas Kit, Agilent) and a precision flow controller (640B, MKS Instruments) to allow for real-time pressure adjustment based on absolute readings of the drift tube pressure using a capacitance manometer (CDG 500, Agilent). For DTIMS measurements, ions were passed through an inlet glass capillary, focused by a high-pressure ion funnel, and accumulated in an ion funnel trap. Ions were then pulsed into the drift tube filled with ~ 3.95 torr of nitrogen gas, where they travelled under the influence of a weak electric field (10-20 V/cm). Ions exiting the drift tube were refocused by a rear ion funnel prior to QTOF MS detection and their arrival times and masses were recorded. All single field IMS-MS data files were populated with collisional cross section (CCS) values using the Agilent IM-MS Browser software and *m/z* and CCS features for each sample were extracted using the Agilent Mass Profiler software.

All samples were run in triplicate and methanol blanks were injected between each sample's replicates to minimize carryover. Thus the 10 mixture samples and 1 DMSO blank resulted in 33 ESI sample runs and 33 ESI blanks, totaling 66 runs in positive mode and another 66 runs in negative mode. This same process was repeated for APPI.

2.2 IMS-MS Data Processing and Feature Extraction. Agilent Tune Mix was analyzed immediately before each set of samples and the associated data was used for mass calibration. Briefly, the first value in the DefaultMassCal.xml (calibration) file within the .D file of the Tune Mix run was multiplied by the square root of the average of the percent differences between expected and measured masses and then replaced within the calibration XML file. This updated XML file was then copy/pasted to all runs immediately after the Tune Mix run. Using Agilent's MassHunter Browser, the Tune Mix run immediately before each set of samples was then used to calibrate all succeeding files using the CCS Calibration tool. A text file named "MfeSettings.txt" was then added to each .D file containing the text "Infusion Yes", a requirement for Agilent's MassProfiler software, which was used to extract and align features using a mass tolerance of 15 ppm + 2 mDa. For the "measure of abundance" setting, "max ion instensity" was selected and "Ion intensity" was set to >= 100 count. The isotope model was unbiased and the charge state was limited to 1-2. There were no mass bounds. Note, no features (with an intensity of more than 100) were downselected or used for building evidence for the presence of molecules using MassProfiler, this was completed later using MAME.

SI 3. Fourier Transform Ion Cyclotron Resonance-Mass Spectrometry.

3.1 Instrumentation. Ultra-high resolution MS experiments were performed by use of a 21-Tesla Fourier transform-ion cyclotron resonator with a Velos Pro dual linear quadrupole mass spectrometer (FTICR-MS) that was designed and constructed in-house at the Environmental Molecular Sciences Laboratory, a national scientific user facility located on the campus of Pacific Norwest National Laboratory.[3] The Velos Pro provides high sensitivity, efficient ion isolation, tandem mass spectrometry (e.g., collision induced dissociation; CID), and automatic gain control (AGC). Ions are transferred from the Velos Pro to the ICR cell for high resolution mass analysis by RF-only quadrupole ion guides. Three stages of differential pumping yield an ultimate pressure of <10^{-10} Torr in the analyzer stage. Signal was acquired using a harmonized ICR cell that utilizes external shimming to approximate an ideal quadrupolar electric field. Using etched fused silica emitters (360 mm OD, 50 mm ID), 100 µL samples were directly infused via ESI at 0.5 mL/min. Methanol blanks were injected between samples to minimize carryover. The electrospray voltage was ±3.0 kV, and the inlet capillary temperature was 300 °C. Mass spectra generated from the average of 200 transient acquisitions by use of an AGC target of 3×10^6 for the *m/z* range ~100–1200.

3.2 FTICR-MS Data Processing and Feature Extraction. An in-house developed script for extracting features detected by the Thermo Xcalibur software (version 2.2) with the Interactive Chemical Information System (ICIS) algorithm was used to generate lists of peaks. Only spectral peaks with a signal-to-noise ratio greater than 2 were used in chemical formula assignment by the Formularity software.[4] Briefly, the Compound Identification Algorithm (CIA) search function in the Formularity software was used to assign chemical formulae consisting of elements C, H, N, O, P, S to peaks according to our previously described method.[5,6] To assign isotopic peaks for compounds consisting of additional elements, we used Formularity's Isotopic Pattern Algorithm (IPA) search function. The IPA database of isotopic peaks for the ~4,700 original ToxCast substances queried by a search function was compiled using the Ecipex software.[7] Once high-mass resolution data was collected using FTICR-MS, there was evidence for a significant presence of chlorinated compounds, leading to the additional calculation of chlorinated isotopic signatures (giving a total of four adducts per molecule with calculated isotopic signatures). Peaks with a signal-to-noise ratio ≥2 were searched within 0.5 ppm mass measurement accuracy. Only chemical formulae with at least 3 major peaks and a d2 fit score (a score based on the Euclidean distance between observed isotopic pattern and expected isotopic pattern) <0.5 were reported as detected.[4] Note, Formularity was not used on the IMS-MS data sets due to instrumental error being too high to reliably assign formulae to potential isotopic signatures.

SI 4. Collisional Cross Section (CCS) Calculations.

We recently developed an automated high-accuracy method for calculating CCS and other chemical properties called the *in silico* Chemical Library Engine (ISiCLE), only requiring chemical structure information (e.g., as provided by the InChI or SMILES[46])

The ISiCLE module for calculating IMS CCS for molecules has three methods for calculating CCS — *Standard*, *Lite*, and *ab initio molecular dynamics (AIMD)-based* — of which the *Standard* and *Lite* methods were used for this study. Complete details regarding CCS calculation methods are provided in SI 4.0. At its current stage of development, the *Standard* method has an average error of 3.2% and the *Lite* method has an average error of 6.7% (Table S3);[47] however, the *Lite* method is much less computationally intensive, making it more than 200 times faster. The *Standard* method was used for calculating the CCS of all three adducts from a selected subset of 1,000 molecules that showed significant evidence (early in the analysis) of being present in the mixtures. The *Lite* method was then used for the remaining molecules, as an appropriate tradeoff between accuracy and computational cost based on the scope of the project. The CCS calculation method and results for each entry in the suspect library are provided in the Supplemental Data. Details on how the ToxCast library was processed to generate the suspect library are provided in SI 5.1.

4.1 Adduct Generation. InChIs[8] from the ToxCast were converted to their "processed InChI" format. Molecular salts were first desalted within ChemAxon's MarvinSketch (chemaxon.com) and exported back to an InChI. InChIs were then neutralized by removing the charge layer within each InChI. If this caused the InChI to be invalid, the processed InChI for that compound was simply left blank and CCS was not calculated for it. The desalted, neutralized InChIs were then converted to their major tautomer, identified using ChemAxon's cxcalc module (v 16.11). These processed InChIs were used to calculate the mass, formula, and CCS of the compound. The formula was first extracted from the InChI using ChemAxon's cxcalc module (v16.11) and the mass was calculated from this formula using molmass.py, a Python package created by Christoph Gohlke (www.lfd.uci.edu/~gohlke/).

Neutral 2D structures were then generated from the processed InChIs using OpenBabel v 2.4.[9] The tautomers were identified using ChemAxon's Calculator, cxcalc (v 16.11). The 2D structure was converted into an initial 3D structure using the general Assisted Model Building with Energy Refinement (AMBER) force field (GAFF) in OpenBabel.[10] Ionization sites in the 3D structure were identified based on pKa values (calculated using cxcalc) to create protonated, sodiated, and deprotonated structures of each compound.

4.2 Lite CCS Calculation Method. CCS was calculated for all adducts (protonated, sodiated, and deprotonated) using the Ion Mobility Projection Approximation Calculation Tool (IMPACT)[11] with the following settings: take hydrogens into account, 64 shots per rotation, 0.001 convergence threshold, and 64 independent runs. Since the calculated CCS are for helium gas (CCS_{He}) , the following equation is used to convert to CCS as measured in nitrogen gas (CCS_{N2}):

$$CCS_{N2} = CCS_{He} + \alpha m^{\beta} \qquad (1)$$

Where m is the mass of the parent compound (i.e., the adduct's mass is not considered). α and β were found by fitting CCS_{He} output from IMPACT to experimental CCS_{N2} values in our validation set[12] using least squares minimization, resulting in values of 27.9 and 0.14, respectively.

4.3 Standard CCS Calculation Method. CCS values were predicted for a subset of the ToxCast library as follows, as described previously.[13] Starting with the generated adducts, molecular dynamics simulations with ten simulated annealing steps were performed using AMBER[14] (v. 14) to generate conformers of each ionized structure at 300 K. Ten conformers from each simulated annealing step were randomly selected and then downselected to three by identifying the two most dissimilar conformers among the ten and the one most similar conformer to the other nine. Finally, a total of 30 conformer geometries were selected for further geometry optimization with density functional theory (DFT) using NWChem.[15] The N_2 CCS value for every DFT-optimized conformer was calculated using the trajectory method implemented in a modified version of MOBCAL.[16-18] Boltzmann weighting based on DFT energies was used to shift the overall CCS distribution toward high probability conformers, thus creating CCS distributions that are characteristic of IMS. DSSTox Structure Identifier (DTXCID)s that had CCS values calculated using the *Standard* vs. *Lite* methods are provided in the Supporting Data.

SI 5. Feature Matching and Scoring using the Multi-Attribute Matching Engine (MAME).

5.1 Library Processing. The library provided by the EPA contained 4,737 substances. This was reduced to 4,348 compounds by (i) removing compounds with a mass below 50 Da or above 1000 Da, (ii) removing compounds with processed InChIs that could not be neutralized without creating an invalid compound, and (iii) combining compounds with the same processed InChI. Because of this last step, our analysis and statistics were based on assessing evidence for which structures may be present, not on determining the correct parent compound. For example, cyclohexylamine and cyclohexylamine hydrochloride were both suspects provided in the ToxCast library. Since these structures are indistinguishable in solution due to the loss of the salt, they were grouped into a single entry within our suspect library. We define suspected presence of these compounds to mean we report both parent compounds as potential candidates when one or both was spiked into a mixture. Note, isomers are not grouped into the same library entry since they do have different structures in solution and, considering the method we describe here, they can be teased apart using CCS (Figure S10).

5.2 Feature Matching. Python (v 2.7.10)[19] was used with the following packages: numpy (v 1.9.3),[20] matplotlib (v 1.5.0),[21] scipy (v 0.16.1),[20,22] and openpyxl (v 2.3.5).[23]

The downselected list of IMS-MS analytical features were used to compare to possible matches within the ToxCast library. If a feature had the same mass as an entry in the suspect library (within the ±6 ppm allowed error), it was scored (as described below) and then recorded as a feature match. This was repeated for all entries and features. Note, CCS was not used at this stage. Any features with matching mass could be matched to a given library entry. CCS was only used to add *additional* evidence.

FTICR-MS features (characterized by a highly accurate measured mass) were evaluated in a similar manner to IMS-MS features but there was only a single run for each mix and the intensity cutoff was set to 1. Because of this, features were removed if they were seen in the blank run at any intensity. When matching to the library, an allowable mass error of ±1.5 ppm was used. These features were also, independently, used by Formularity (with a mass error of ±0.5 ppm) to match chemical formulas in the ToxCast library to their isotopic signatures.

5.3 Scoring. Once all analytical features were processed and matched to corresponding entries in the ToxCast, we scored the evidence of each ToxCast entry being in each mixture using the following weighting method (indices correspond to Table 2):

For IMS-MS, features matched to a library entry can accrue points contributing to the total evidence score based on the following scenarios. If the average intensity of that feature across all replicates is more than the 30th percentile value of all passing feature intensities (within that mode), then it would score 2.0 points (index 1). Otherwise, if it fell below this cutoff, it would score 1.0 point (index 2) since lower intensities could possibly be caused by noise or low-level contamination. An additional 3.0 points (index 3) could also be earned by each feature demonstrating a measured vs. calculated CCS error ≤ ±5%.

For FTICR-MS, features matched to a library entry can accrue points contributing to the total evidence score based on the following scenarios. If the average intensity of that feature across all replicates is more than the 30th percentile value of all passing feature intensities (within that mode), then it would score 4.0 points (index 4). Otherwise, if it fell below this cutoff, it would score 2.0 points (index 5). An additional 3.0 additional points (index 6) could be earned if the predicted isotopic distribution of its chemical formula matched to a corresponding experimental isotopic distribution, found using Formularity, as described in SI 3.2. These 3.0 points could only be earned once per library entry since they are not specifically tied to a given feature.

For both IMS-MS and FTICR-MS, for each adduct, an additional 0.5 points (index 8) was added for each feature after the first matched feature, meaning each of these features lends their own evidence to the total evidence score but also the fact there is more than one matched feature to a given adduct further increases evidence that adduct was truly present . IMS-MS and FTICR-MS features did not contribute toward each other's counts. An additional 1.0 point (index 7) was also given for each adduct observed after the first match (again, IMS-MS and FTICR-MS did not contribute toward each other's counts). For example, if features corresponding to the $[M+H]^+$ and $[M+Na]^+$ of a given library entry are found, additional evidence is added since there is evidence of more than one adduct (whether or not these adducts are from the same mode). If at least one IMS-MS and one FTICR-MS feature was matched to a library entry, an additional 2.0 points (index 9) was added to the total score. Next, there were two final sources of points: 4.0 points (index 10) if the matched library entry had a unique mass (closest neighbor by exact mass more than 6 ppm away) and 1.0 point (index 11) if the matched library entry had an exact mass ≥ 200 Da (due to lower mass errors frequently seen for larger compounds). Finally, any points collected due to a mass match (all points except those obtained from a low error CCS match) were divided by the number of other library entries within a ±6 ppm mass range. For example, if two compounds in the ToxCast library had the same mass, their points from everything except CCS would be divided by two, lowering their overall evidence scores. As an example, Figure 2 shows how pioglitazone was scored and correctly labeled as suspected present in one of the mixtures. At the end, for each mix, if a library entry had accumulated a score of 6.0 or greater, it was then labeled as "suspected present" in the mix and assigned an evidence level, as described in the following section.

5.4 Metrics. As metrics to quantify success, we used false discovery rate (FDR, the percentage of false positives out of the total number of compounds labeled as suspected present), false negative rate (FNR, also known as the miss rate, the percentage of false negatives out of how many compounds were spiked in by the EPA), and accuracy (the percentage of correct labels). Equations for each of these metrics are provided below. The overall goal of the method is to minimize FDR and FNR, while maximizing overall accuracy. When reporting these values here, we use the average across the ten mixtures. These metrics (and more) are broken down for each mixture in the Supplemental Data.

It is important to note the analysis and statistics were based on assessing evidence for which structures may be present, not on determining the correct parent compound. For example, cyclohexylamine and cyclohexylamine hydrochloride were both suspects provided in the ToxCast library. Since these structures are indistinguishable in solution due to the loss of the salt, they were grouped into a single entry within the suspect library. If there is enough evidence for the desalted compound (cyclohexylamine), we report this grouped entry as suspected present, and consider this a true positive when one or both parent compounds were intentionally spiked in.

Since our approach uses binary classification (i.e., labeling each ToxCast library entry as suspected present or not),[24] we relied on the Equations 1-3 to quantify success.

$$\text{False Discovery Rate (FDR)} = \text{FP} / (\text{FP} + \text{TP}) \qquad (2)$$

$$\text{False Negative Rate (FNR)} = \text{FN} / (\text{FN} + \text{TP}) \qquad (3)$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN}) \qquad (4)$$

Where TP represents the number of true positives (compounds labeled as suspected present and actually were spiked in by the EPA), FP represents the number of false positives (compounds labeled as suspected present but not spiked in), TN represents the number of true negatives (compounds labeled as not suspected present and were not spiked in), and FN represents the number of false negatives (compounds labeled as not suspected present but were spiked in).

Since each of the above equations require a set cutoff (to differentiate between "suspected present" and "not suspected present" labels) to quantify the overall success, we also used the area under the precision-recall curve (AUPR). The precision-recall curve takes into account all cutoffs necessary to range from a precision of 0% (which maximizes recall) to a recall of 0% (which maximizes precision). This helped us to directly compare scoring methods and ultimately identify the most successful, corresponding to the highest AUPR. The precision-recall curve was also used for selecting the optimal cutoff, which is discussed in more detail in the next section.

5.5 Score Optimization. After unblinding, we used Monte Carlo[49] and particle swarm optimization (PSO, via PySwarm[50]) methods, implemented in Python scripts, to select new weights for each scoring criteria using an objective function to maximize the area under the precision-recall curve (AUPR). AUPR is generated by determining precision and recall, which can be derived directly from FDR and FNR, respectively, parameterized by a minimum evidence score threshold. This enables performance of the scoring weights to be assessed without an explicit threshold selection for the suspected presence score, which is a nontrivial decision with implications beyond the scope of this work.[51] AUPR was also selected as the objective function due to its relatively good performance compared to other classifiers when dealing with imbalanced datasets (i.e., significantly more true negatives compared to true positives).[52]

In total, there are eleven adjustable weights when scoring a compound, each of which are described in the Scoring section (SI 5.3). Two methods were used to adjust weights: Monte Carlo[25] and particle swarm optimization (PSO, via PySwarm).[26] Both methods allowed weights between 0 and 10. The set goal for both methods in our case was maximizing the area under the precision-recall curve (AUPR), calculated using sklearn's metric package.[27] The equations for precision and recall are given in Equations 4-5.

$$\text{Precision} = \text{TP} / (\text{FP} + \text{TP}) = 1 - \text{FDR} \qquad (5)$$

$$\text{Recall} = \text{TP} / (\text{FN} + \text{TP}) = 1 - \text{FNR} \qquad (6)$$

Three separate Monte Carlo simulations were run for 100,000 iterations each. Three PSO simulations were run with the following parameters: *swarmsize=3000, omega=0.5, phip=0.5, phig=0.5, maxiter=500, minstep=0.001, minfunc=1e-8, debug=False, processes=1, particle_output=False*.

To find the new optimal evidence score threshold, and consequently the low evidence threshold, we used the threshold that yielded the maximum F1 score (Eq. 6). The medium and high score thresholds were then chosen to accomplish an FDR of 50% and 10% (when considering Level 2a matches only), respectively.

$$\text{F1 Score} = 2^* (1\text{-FDR}) * (1\text{-FNR}) / ((1\text{-FDR}) + (1\text{-FNR})) \qquad (7)$$

A "brute force" Monte Carlo run was also performed by testing a value of 0.0, 5.0, and 10.0 for each of the 11 weights. This lead to a total of 177,147 combinations (eleven weights, testing three values each). Results shown in Figure S12.

5.6 Sensitivity Analysis. The weights found using Monte Carlo and Pyswarm were tested for their overall sensitivity. To do this, each of the weights were tested using a combination of optimized values, optimized values + 0.5, and optimized values – 0.5. This led to a total of 177,147 combinations (eleven weights, testing three values each).

Results

SI 6. Experimental Analysis of Samples and Feature Extraction.

A total of 14 data sets were successfully generated per sample (plus additional data for blanks). The raw IMS-MS data included ~200,000 total *m/z*-CCS features observed across triplicate analyses and ~450,000 total *m/z* features observed across FTICR-MS analyses (Figures S2-9). This data showed evidence of many different adducts, as indicated by the high number of features and a significant presence of multimers, as indicated by frequent observations of features with extremely high CCS: *m/z* ratios (Figures S4-5 and S7-8, and see Figure S1b for the much tighter *m/z*-CCS distribution of the library when considering [M+H] + features).[37] In future analyses of similar samples, appropriate dilution will be necessary to reduce the formation of multimers. To help reduce noise and low-level contamination, we downselected to a subset of features using constraints based on feature intensity and presence across technical replicates (and absence in blanks), applying the cutoffs described in Table 1.

For IMS-MS, an intensity cutoff of 1,000 was set for all features in addition to requiring that each feature must have been observed across all three technical replicates and no more than once across the corresponding blank replicates. This removed 94% of features. There was still significant evidence of multimer formation in the positive mode ESI-IMS-MS analyses (Figure S4b) but, with no way to ensure these were removed without losing possible overlapping monomer features, we decided to move forward, understanding that most of the suspected multimer features would not match the CCS values of library entries. For FTICR-MS, we did not find a reliable method to apply an intensity cutoff, so the intensity cutoff was trivially set to 1. Since there was only a single replicate for each condition (due to limited sample), any feature seen at any intensity in the blank was removed, leading to 5% of features being removed.

SI 7. Scoring Optimization.

Both optimization methods were able to increase the AUPR compared to the original weighting scheme, which had an AUPR of 0.35. PSO achieved the highest AUPR of 0.41 (Figures S13-14), while the Monte Carlo method achieved an AUPR of 0.39. To further investigate these optimized weights, we performed a sensitivity analysis (Figure S15, description provided in SI 5.6). We found the top PSO result to be much less stable than the top Monte Carlo result. It is possible that the PSO result may have been overfit to this particular dataset. Due to this finding, and the need to use a stable, general scoring system that may work well with real complex samples, we selected the optimized weights returned by the best Monte Carlo run to be the final set (shown in Table S2). All 4 scoring methods discussed here are also represented by confusion matrices in Figure S16. Applying these values to reassess evidence and using a score cutoff of 11.2 (the cutoff that yielded the highest

F1 score), in respect to our blinded method, we were able to decrease FDR by 24% (on absolute terms, from 77% to 53%) and increase accuracy by 5% (to 96%). Consequently, FNR increased by 10% (to 67%) due to the tradeoff between FNR and FDR, but note the magnitude of this increase is lower than the magnitude of the FDR decrease. For molecules suspected to be present with high evidence scores (49.0 or more) provided by both mass and CCS, the FDR dropped from 35% to 10%. Similar to our results prior to scoring optimization, FDR trended smoothly with the magnitude of evidence score (Figure S14f). Furthermore, the number of compounds labeled as suspected present per sample dropped from an average of 354 to 127 (the mixtures had an average of 194 substances per sample).

As mentioned in the main manuscript, of the eleven adjustable point-awarding criteria, four were consistently awarded the highest weight by both optimization methods (in order from the highest): (i) multiple adducts being observed by a single instrument (index 7), (ii) high intensity IMS-MS features (index 1), (iii) low intensity IMS-MS features (index 2), and (iv) detection on both MS instruments (index 9) (Figure S13). These were deemed to be of very high importance for determining evidence. The lowest scoring criteria were (in order from lowest): (i) high intensity FTICR-MS features (index 4), (ii) unique mass (index 10), and (iii) large mass (index 11). "Unique mass" (index 10) scoring among the lowest was not expected, but it does reveal that our method's success may be less dependent on the size of the library than expected. A larger library size can lead to fewer unique masses and a higher dependence on multiple matched features. This highlights that even though library entries may be indistinguishable in the mass dimension (i.e., identical chemical formula), it is unlikely that all associated adduct CCS values will overlap.

Regarding the overall low weights assigned to FTICR-MS properties, this could have been due to using an excessively large mass error window for these features (about three times larger than needed) as well as using a cutoff intensity of 1. Decreasing the allowable mass error and increasing the intensity cutoff may help to increase the reliability of matched FTICR-MS features. These are easy values to change within MAME but were left as-is in order to match the results sent to the EPA prior to unblinding. Furthermore, using only singlet samples for the FTICR-MS experiments, instead of triplicate like the IMS-MS experiments, may have resulted in more noise being considered as legitimate features. Note, though, that with "Detection by both MS's" (index 9) weighted highly, FTICR-MS features are still shown to be important, but only when corresponding IMS-MS features are also present.

CCS (index 3) was assigned a weight lower than other IMS-MS features (indices 1-2), but was still weighted more heavily than high intensity FTICR-MS (index 4) and library-related criteria (indices 10-11). Even with our lowest-accuracy CCS calculation method (the Lite method) used for 84% of molecules, CCS was still a useful addition to our multi-attribute approach by contributing evidence to 82%, 91%, and 94% of low, medium and high evidence true positives, respectively.

SupportingData.xlsx Captions

SD 1. Suspect library.

Suspect library. This was derived from the original ToxCast library sent by the EPA using the method described in SI 5.1. Columns with labels Mix{#} (with # ranging from 499 to 508) indicate which compounds were in each mixture. 0 indicates the library entry was not spiked into that mixture and 1 indicates it was.

SD 2. Blinded Scores.

Scores awarded to each library entry using the parameters described before being unblinded (Table 1 & 2 in the main text).

SD 3. Blinded Statistics.

Results based on the scores given in SD 2 and the suspected present vs. not suspected present score cutoff (6 points).

SD 4. Unblinded Scores.

Scores awarded to each library entry using the parameters described before being unblinded (Table 1 & 3 in the main text).

SD 5. Unblinded Statistics.

Results based on the scores given in SD 4 and the suspected present vs. not suspected present score cutoff (11.2 points).

Tables

Category	Parameter	Cutoff
IMS-MS	Intensity	≥ 1000 a.u. [a]
	Mass Error (Magnitude)	≤ ± 6 ppm
	# Seen in Samples	3
	# Seen in Blanks	≤ 1
	CCS Error	≤ 5%
FTICR-MS	Intensity	≥ 1 a.u.
	Mass Error (Magnitude)	≤ ± 1.5 ppm
	# Seen in Samples	1
	# Seen in Blanks	0
IMS-MS & FTICR-MS	High Intensity	≥ 30th %ile
	Low Intensity	< 30th %ile
Library	Unique Mass [b]	> ± 6 ppm
	Large Mass	≥ 200 Da

[a] Arbitrary units. [b] A library entry’s mass is considered unique if its nearest neighbor library entry is more than 6 ppm away.

Table S1. Parameter cutoffs used for scoring and downselection.

Category	Index	Criteria	Weight
IMS-MS	1	High Intensity	9.1
	2	Low Intensity	1.0
	3	Low CCS Error	1.1
FTICR-MS	4	High Intensity	0.2
	5	Low Intensity	2.2
	6	Isotopic Signature	0.6
IMS-MS & FTICR-MS	7	Additional Adducts	9.5
	8	Additional Features	1.1
	9	Detected by Both MS's	2.3
Library	10	Unique Mass	0.2
	11	Large Mass	0.9

Table S2. Optimized weights for scoring criteria.

	Average Unsigned Error		
Method	**Validation Set**	**diCQA[13]**	**This Study***
Lite	6.7%	4.8%	4.8%
Standard	3.2%	2.6%	3.2%
AIMD-based	N/A	0.8%	N/A

*For true positives. AIMD: *ab initio* molecular dynamics; diCQA: dicaffeoylquinic acid.

Table S3. Error distributions for calculated IMS CCS methods within ISiCLE. Please refer to Colby et al.[12] for more details.

Figures

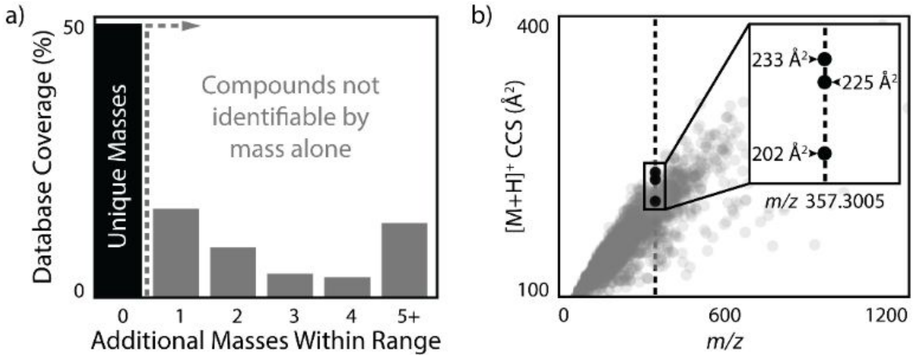


Figure S1. CCS is a chemical property that increases each library entry's uniqueness. (a) Number of molecule entries in the ToxCast library (shown as a percentage of total library size) whose protonated masses fall within ±6 ppm of another entry. Zero (black bar) indicates no neighbors within this mass range (a molecule that can be resolved with mass alone, 2,216 total). The grey bars represent molecules (2,130 total) that cannot be distinguished based on mass alone, within an instrumental error of ±6 ppm. (b) Calculated CCS vs. m/z for the protonated forms of each molecule in the ToxCast library. The inset shows the example of m/z 357.3005, where 3 molecules lie within ±6 ppm of one another. When adding the property of CCS, all 3 molecules are predicted to become analytically unique within our specified parameter thresholds.

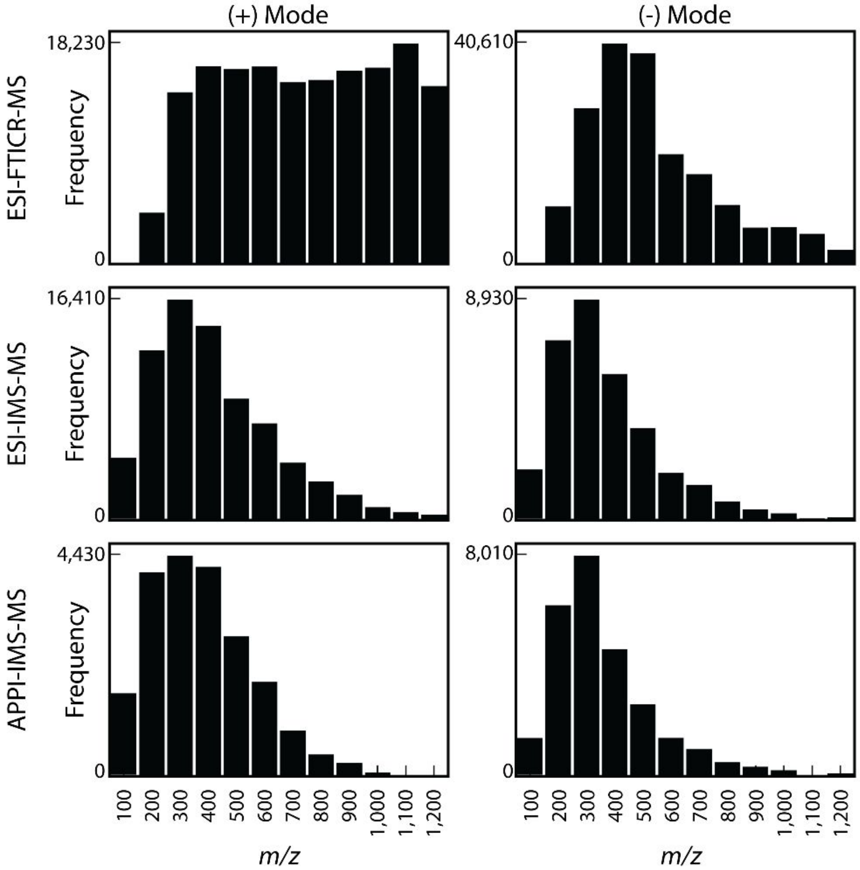


Figure S2. Distribution of experimental features observed with FTICR-MS and IMS-MS.

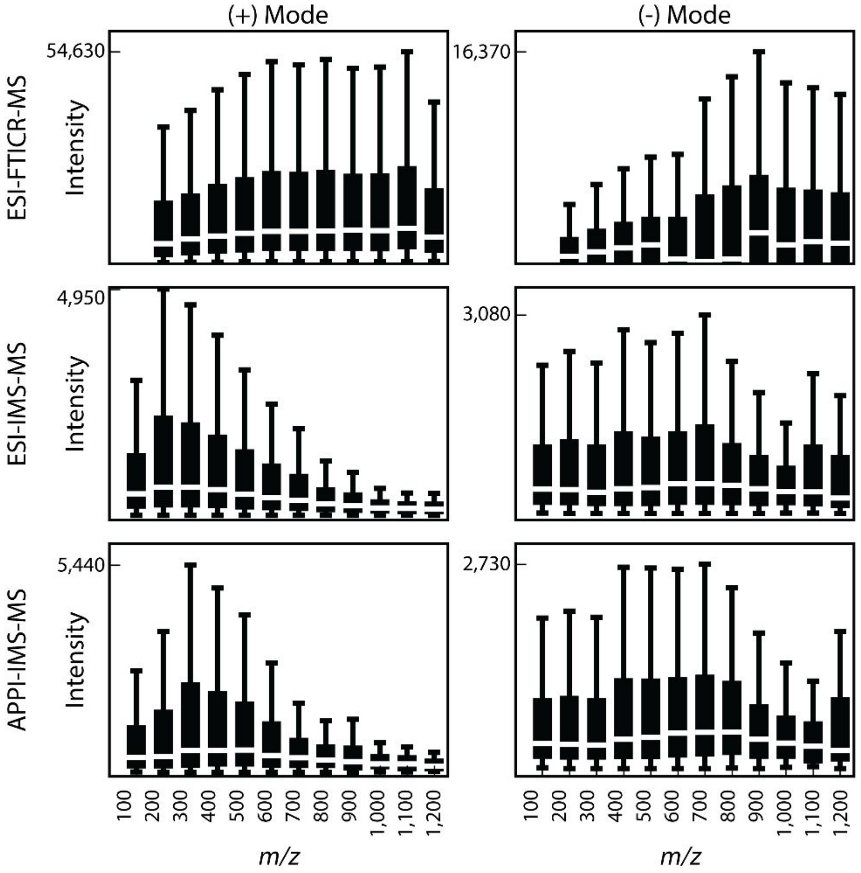


Figure S3. All features masses observed with FTICR-MS and IMS-MS across different mass bins. Bin labels indicate the range as [x-50, x + 50).

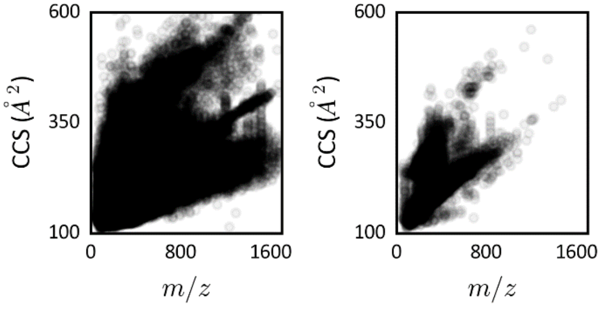


Figure S4. Positive mode ESI-IMS-MS features. (a) All features before selection (n = 123,580). (b) All features that passed the downselection process (n = 8,002). There still appears to be a significant presence of multimers, as shown by two distinct trends in CCS:mass (the trend with the higher CCS:mass likely indicating multimers). Note, many of these features overlap with APPI features due to feature alignment using MassProfiler (see SI 2.2 for more information).

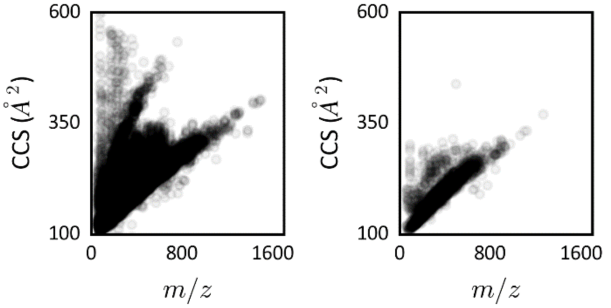


Figure S5. Positive mode APPI-IMS-MS features. (a) All features before selection (n =21,036). (b) All features that passed the downselection process (n =2,539). Note, many of these features overlap with ESI features due to feature alignment using MassProfiler (see SI 2.2 for more information).

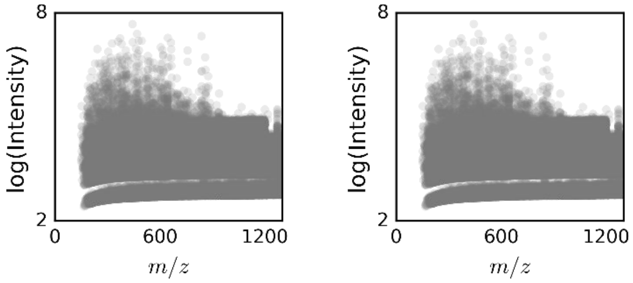


Figure S6. Positive mode ESI-FTICR-MS features. (a) All features before selection (n =258,059). (b) All features that passed the downselection process (n =250,013).

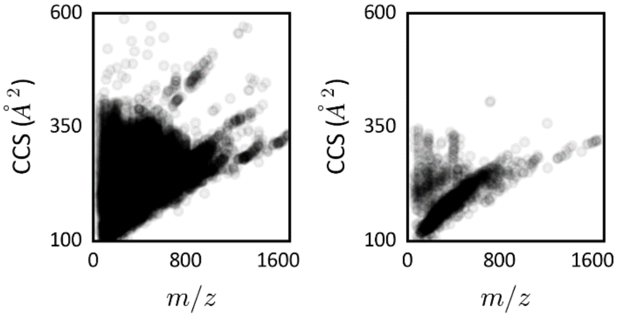


Figure S7. Negative mode ESI-IMS-MS features. (a) All features before selection (n =21,531). (b) All features that passed the downselection process (n =1,782). Note, many of these features overlap with APPI features due to feature alignment using MassProfiler (see SI 2.2 for more information).

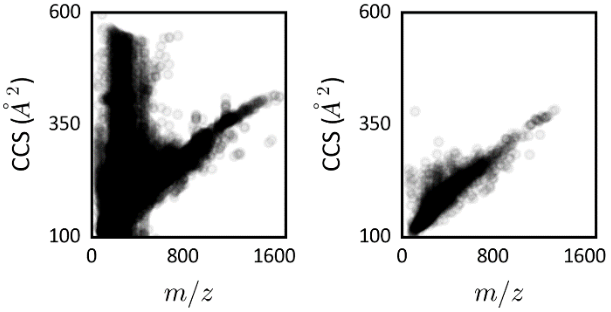


Figure S8. Negative mode APPI-IMS-MS features. (a) All features before selection (n =26,964). (b) All features that passed the downselection process (n =2,483). Note, many of these features overlap with ESI features due to feature alignment using MassProfiler (see SI 2.2 for more information).

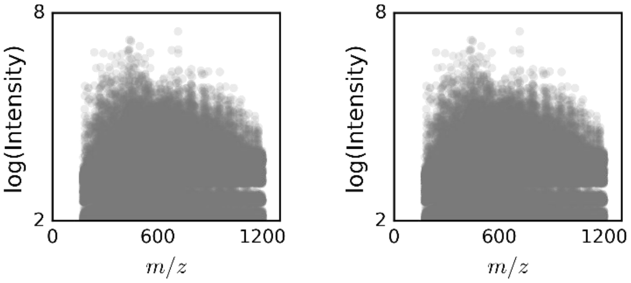
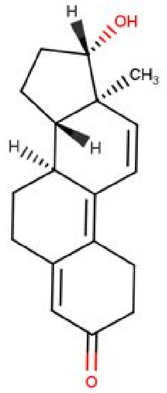
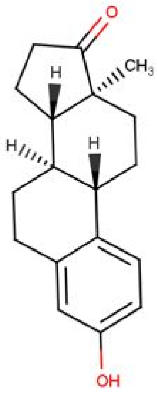


Figure S9. Negative mode ESI-FTICR-MS features. (a) All features before selection (n =189,588). (b) All features that passed the downselection process (n =185,139).

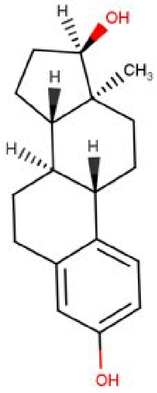
17β-Trenbolone
$C_{18}H_{22}O_2$
Confidence Score: 7.4
True Positive



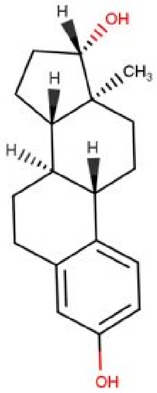
Estrone
$C_{18}H_{22}O_2$
Confidence Score: 4.4
True Negative



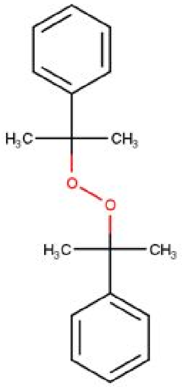
17α-Estradiol
$C_{18}H_{24}O_2$
Confidence Score: 0
True Negative



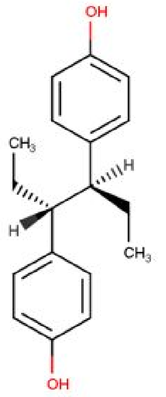
17β-Estradiol
$C_{18}H_{24}O_2$
Confidence Score: 0
True Negative



Dicumyl peroxide
$C_{18}H_{22}O_2$
Confidence Score: 4.4
True Negative



meso-Hexestrol
$C_{18}H_{22}O_2$
Confidence Score: 4.4
True Negative



4,4'-(4-Methylpentane-2,2-diyl)diphenol
$C_{18}H_{22}O_2$
Confidence Score: 4.4
True Negative

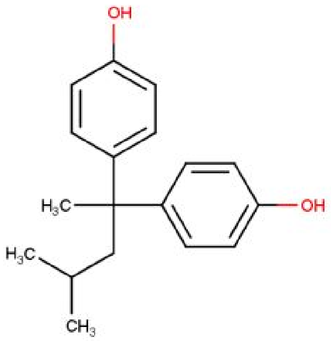


Figure S10. Evidence scores and results for 17β-Trenbolone (a compound spiked into the second mixture) and the 6 other compounds in our processed ToxCast library with the same formula ($C_{18}H_{22}O_4$) or same carbon bonding order (carbon layer in the InChI equal to "c1-18-9-8-14-13-5-3-12(19)10-11(13)2-4-15(14)16(18)6-7-17(18)20"). Note: 17β-Trenbolone was the only compound here that was spiked in and was also the only compound labeled as "present" due to an evidence score above 6.0, meaning all assigned IDs here were correct. All scores are given for the second mix.

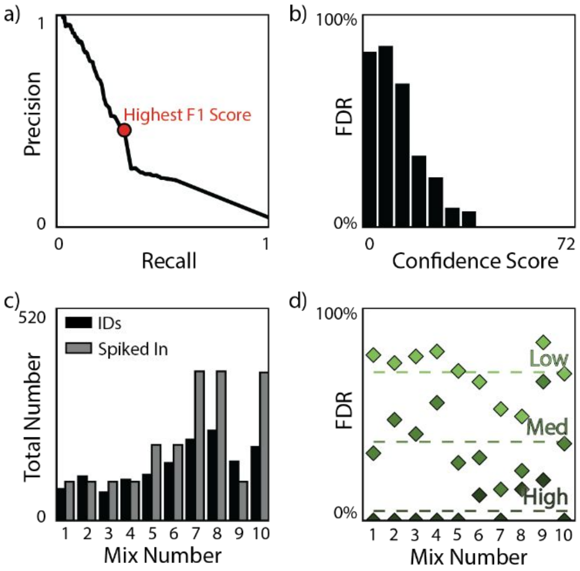


Figure S11. Results using our multi-attribute matching methods, keeping the original set of blinded weights but optimizing the cutoff. (a) AUPR curve, with red dot showing our threshold (a total evidence score of 9.5). (b) FDR as a function of evidence score. (c) Comparison between the number of molecules suspected present compared to the number of molecules spiked into each mixture. (d) FDR for each of the mixes individually, split by evidence levels.

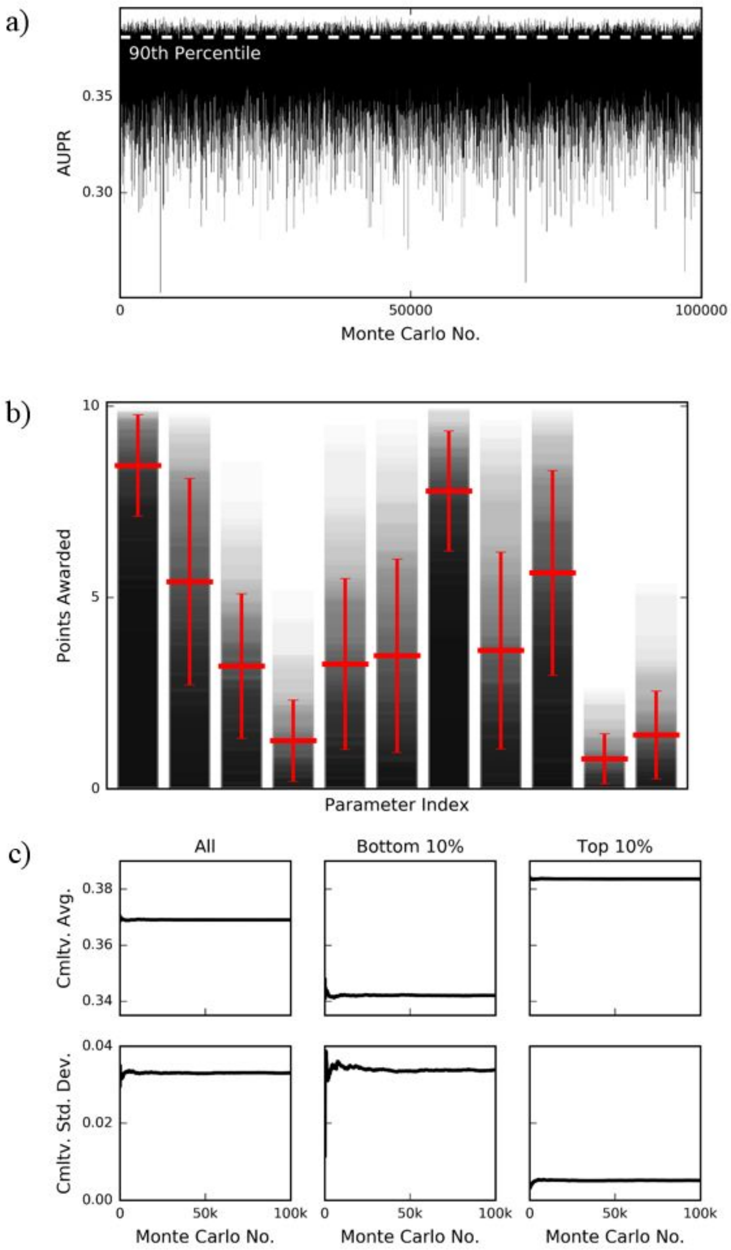


Figure S12. Monte Carlo Results. (a) AUPR score for each Monte Carlo run. (b) Optimal weight distribution using the top 100 scores from the Monte Carlo runs. Red horizontal lines show the average, red vertical lines show the standard deviation. (c) Cumulative average and standard deviation of AUPR across runs.

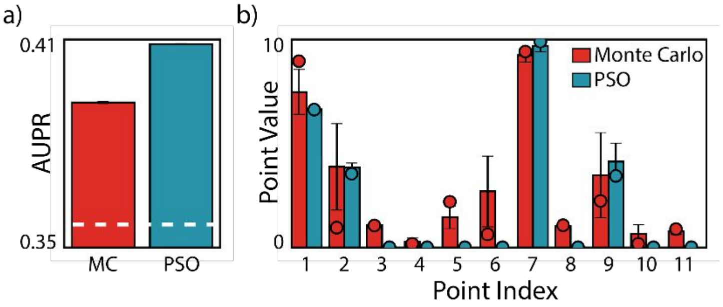


Figure S13. Comparison of Monte Carlo and PSO simulation results for each of the three Monte Carlo simulations and Pyswarm simulations. (a) Average AUPR. Error bars included but their magnitude is close to zero. The white dashed line represents the AUPR for our blinded results. (b) Weights assigned. The average is represented by bar height and the error bars indicate standard deviation. Scatter points represent the weights of the highest scoring set.

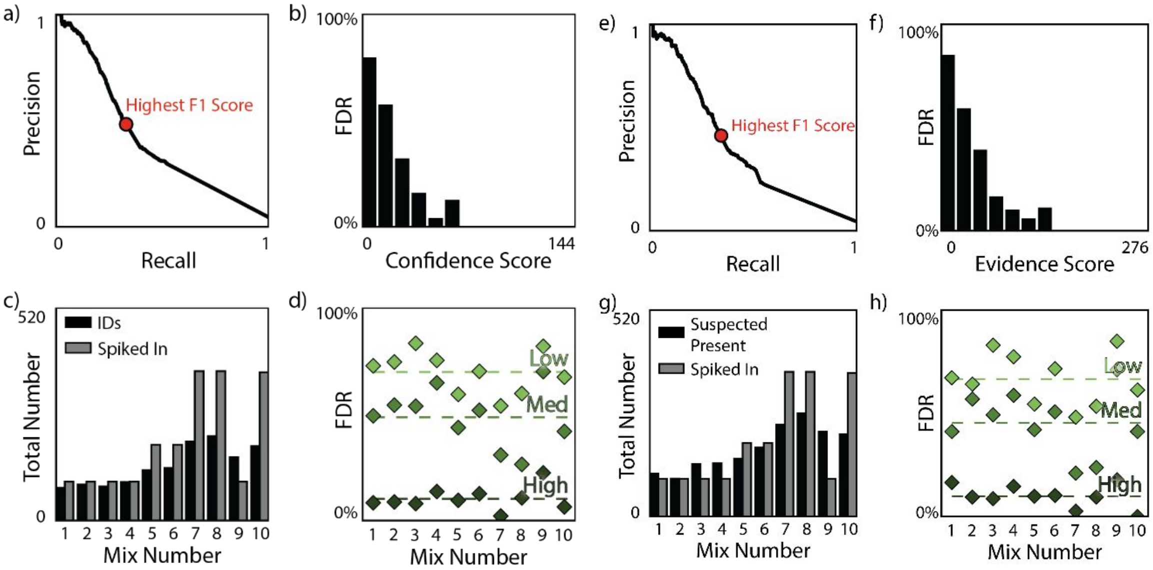


Figure S14. Results using our multi-attribute matching methods, with optimized weights found using (a-d) Optimized, weights chosen using Pyswarm results. (e-h) Optimized, weights chosen using Monte Carlo results. (a, e) AUPR curve, with red dot showing our threshold (a total evidence score of 6.7 and 11.2 for Pyswarm and Monte Carlo optimized results, respectively). (b, f) FDR as a function of evidence score. (c, g) Comparison between the number of molecules suspected present compared to the number of molecules spiked into each mixture. (d, h) FDR for each of the mixes individually, split by evidence levels.

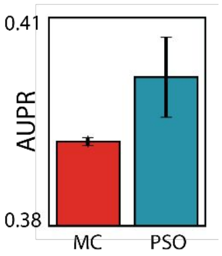
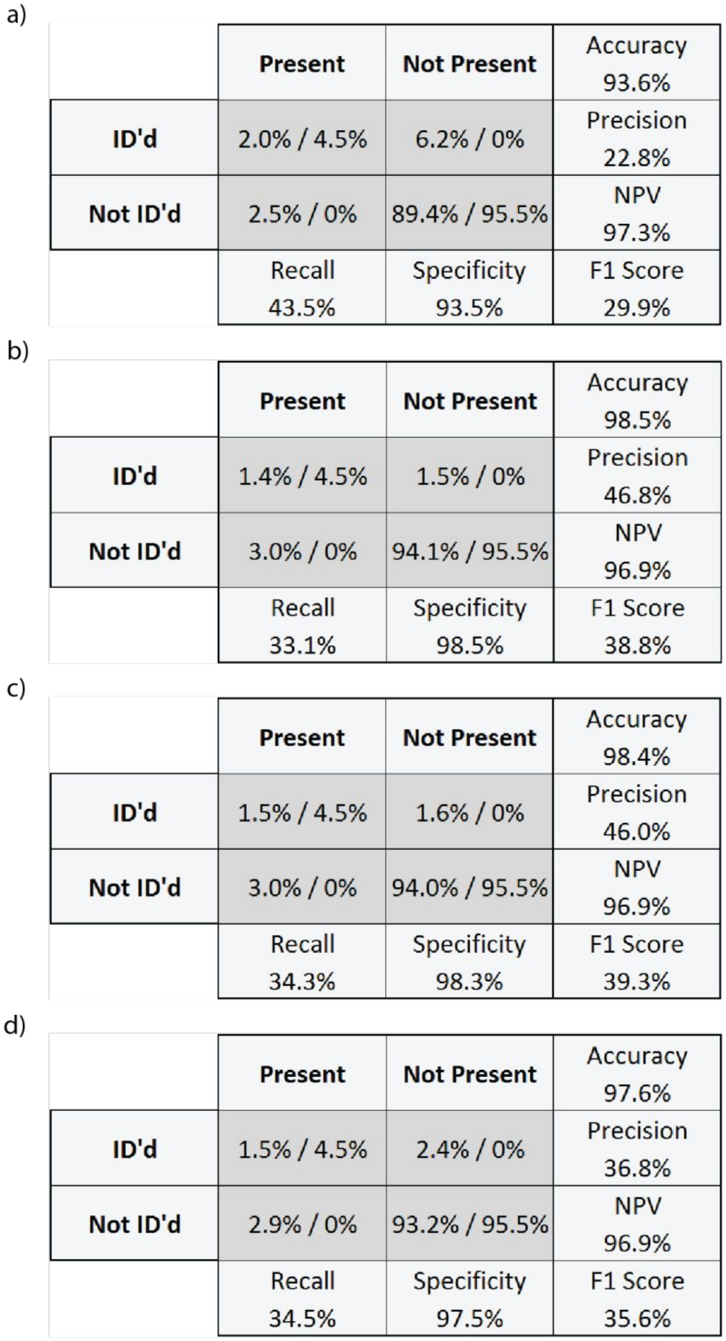


Figure S15. Sensitivity analysis. "MC" stands for Monte Carlo.



a)

	Present	Not Present	Accuracy 93.6%
ID'd	2.0% / 4.5%	6.2% / 0%	Precision 22.8%
Not ID'd	2.5% / 0%	89.4% / 95.5%	NPV 97.3%
	Recall 43.5%	Specificity 93.5%	F1 Score 29.9%

b)

	Present	Not Present	Accuracy 98.5%
ID'd	1.4% / 4.5%	1.5% / 0%	Precision 46.8%
Not ID'd	3.0% / 0%	94.1% / 95.5%	NPV 96.9%
	Recall 33.1%	Specificity 98.5%	F1 Score 38.8%

c)

	Present	Not Present	Accuracy 98.4%
ID'd	1.5% / 4.5%	1.6% / 0%	Precision 46.0%
Not ID'd	3.0% / 0%	94.0% / 95.5%	NPV 96.9%
	Recall 34.3%	Specificity 98.3%	F1 Score 39.3%

d)

	Present	Not Present	Accuracy 97.6%
ID'd	1.5% / 4.5%	2.4% / 0%	Precision 36.8%
Not ID'd	2.9% / 0%	93.2% / 95.5%	NPV 96.9%
	Recall 34.5%	Specificity 97.5%	F1 Score 35.6%

Figure S16. Confusion matrices, demonstrating the overall statistics across all mixtures. (a) blinded results, (b) unblinded results, optimized using Monte Carlo, (c) unblinded results, optimized using PSO, and (d) unblinded results, using the original set of unblinded weights but optimizing the cutoff. NPV represents negative predictive value (TN / (TN + FN)) and specificity represents the true negative rate (TN / (TN + FP)). The format X% / Y% represents the actual performance of the method (X) and the optimal performance (Y). Optimal performance is calculated based on the number of spiked in compounds (possible true positives) and compounds in the library but not spiked in (possible true negatives) and adds up to 100%. False positives and false negatives are set at 0% to represent making no errors. The equations for accuracy, precision, recall, and F1 Score are given in SI 5.4-5.5.

References

(1) May, J. C.; Goodwin, C. R.; Lareau, N. M.; Leaptrot, K. L.; Morris, C. B.; Kurulugama, R. T.; Mordehai, A.; Klein, C.; Barry, W.; Darland, E.; Overney, G.; Imatani, K.; Stafford, G. C.; Fjeldsted, J. C.; McLean, J. A. *Analytical Chemistry* **2014**, *86*, 2107-2116.

(2) Ibrahim, Y. M.; Baker, E. S.; Danielson Iii, W. F.; Norheim, R. V.; Prior, D. C.; Anderson, G. A.; Belov, M. E.; Smith, R. D. *International Journal of Mass Spectrometry* **2015**, *377*, 655-662.

(3) Shaw, J. B.; Lin, T.-Y.; Leach, F. E.; Tolmachev, A. V.; Tolić, N.; Robinson, E. W.; Koppenaal, D. W.; Paša-Tolić, L. *Journal of The American Society for Mass Spectrometry* **2016**, *27*, 1929-1936.

(4) Tolic, N.; Liu, Y.; Liyu, A.; Shen, Y.; Tfaily, M. M.; Kujawinski, E. B.; Longnecker, K.; Kuo, L. J.; Robinson, E. W.; Pasa-Tolic, L.; Hess, N. J. *Anal. Chem.* **2017**, *89*, 12659-12665.

(5) Tfaily, M. M.; Chu, R. K.; Tolić, N.; Roscioli, K. M.; Anderton, C. R.; Paša-Tolić, L.; Robinson, E. W.; Hess, N. J. *Analytical Chemistry* **2015**, *87*, 5206-5215.

(6) Tfaily, M. M.; Chu, R. K.; Toyoda, J.; Tolić, N.; Robinson, E. W.; Paša-Tolić, L.; Hess, N. J. *Analytica Chimica Acta* **2017**, *972*, 54-61.

(7) Ipsen, A. *Analytical Chemistry* **2014**, *86*, 5316-5322.

(8) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *Journal of Cheminformatics* **2015**, *7*, 23.

(9) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *Journal of Cheminformatics* **2011**, *3*, 33.

(10) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of computational chemistry* **2004**, *25*, 1157-1174.

(11) Marklund, Erik G.; Degiacomi, Matteo T.; Robinson, Carol V.; Baldwin, Andrew J.; Benesch, Justin L. P. *Structure* **2015**, *23*, 791-799.

(12) Colby, S. M.; Thomas, D. G.; Nunez, J.; Baxter, D.; Glaesemann, K.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J.; Metz, T. O.; Renslow, R. S. *Analytical Chemistry. In press (10.1021/acs.analchem.8b04567)* **2019**.

(13) Zheng, X.; Renslow, R. S.; Makola, M. M.; Webb, I. K.; Deng, L.; Thomas, D. G.; Govind, N.; Ibrahim, Y. M.; Kabanda, M. M.; Dubery, I. A.; Heyman, H. M.; Smith, R. D.; Madala, N. E.; Baker, E. S. *J. Phys. Chem. Lett.* **2017**, *8*, 1381-1388.

(14) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *Journal of computational chemistry* **2005**, *26*, 1668-1688.

(15) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. *Computer Physics Communications* **2010**, *181*, 1477-1489.

(16) Campuzano, I.; Bush, M. F.; Robinson, C. V.; Beaumont, C.; Richardson, K.; Kim, H.; Kim, H. I. *Analytical Chemistry* **2012**, *84*, 1026-1033.

(17) Mesleh, M. F.; Hunter, J. M.; Shvartsburg, A. A.; Schatz, G. C.; Jarrold, M. F. *The Journal of Physical Chemistry* **1996**, *100*, 16082-16086.

(18) Shvartsburg, A. A.; Jarrold, M. F. *Chemical Physics Letters* **1996**, *261*, 86-91.

(19) python.org.

(20) Walt, S. v. d.; Colbert, S. C.; Varoquaux, G. *Computing in Science & Engineering* **2011**, *13*, 22-30.

(21) Hunter, J. D. *Computing in Science & Engineering* **2007**, *9*, 90-95.

(22) Jones, E.; Oliphant, E.; Peterson, P.; Others. **2001**.

(23) bitbucket.org/openpyxl/openpyxl/src.

(24) Broadhurst, D. I.; Kell, D. B. *Metabolomics* **2006**, *2*, 171-196.

(25) Mooney, C. Z. *Monte Carlo Simulation*; SAGE Publications, 1997.

(26) github.com/tisimst/pyswarm.
(27) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. *J. Mach. Learn. Res.* **2011**, *12*, 2825-2830.